Ecological Informatics 51 (2019) 82–95

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Documents as data: A content analysis and topic modeling approach for analyzing responses to ecological disturbances

Mark Altaweel[a,*], Christopher Bone[b], Jesse Abrams[c]

[a] Institute of Archaeology, University College London, UK
[b] Department of Geography, University of Victoria, UK
[c] Warnell School of Forestry and Natural Resources, Savannah River Ecology Laboratory, University of Georgia, USA

ARTICLE INFO

Keywords:
Mountain pine beetle
Topic modeling
Hierarchical Dirichlet process
Latent Dirichlet allocation
Term frequency–inverse document frequency
Content analysis
Natural language processing

ABSTRACT

We apply content analysis on government documents containing ecological information relevant to a significant ecological disturbance - mountain pine beetle (MPB) outbreaks in the United States. The intent is to demonstrate a semi-automated approach that applies topic modeling to investigate policy responses to ecological disturbances, using latent Dirichlet allocation (LDA) and hierarchical Dirichlet process (HDP), and term frequency–inverse document frequency (tf-idf) analysis. Results demonstrate how analysts and researchers are better able to understand what topics and focus areas government officials consider in relation to MPB disturbances. In the case study demonstrating the method's utility, documents found from before 1960 and until recent years demonstrate focus on outbreak area, tree mortality, research and services, management, infestation, outbreak control, fire, insect control, outbreak factors, and tree population. Terms such as *fire*, *mortality*, *treatment*, and *outbreak* reflect more recent U.S. government focus on MPB, while *disease* and *infestation* have become less of a focus in recent years. There are also varying differences and interests between how different parts (i.e., federal agencies versus congress) of the U.S. government focus on MPB, where mostly interests and focus are not aligned or do not match temporally. As a term, *temperature* has become a greater recent government focus, but there is general avoidance of the term *climate change*. The methods applied demonstrate the utility of topic modeling and tf-idf for understanding discourse and content in policy related to ecological disturbances. The tool created in this effort is provided freely as a way for scientists and researchers to extend its utility in ecological policy research.

1. Introduction

Ecological disturbances influence how ecosystems function and evolve over time by creating changes in local conditions that potentially lead to larger-scale impacts. Disturbances such as wildfires, flooding, and windstorms can occur quickly, causing abrupt shifts in ecosystem processes that endure for several years (Thom et al., 2013). Slower-moving disturbances, such as outbreaks of insects that kill or weaken host organisms, can occur over longer periods and exert more gradual ecosystem impacts that influence ecological processes for decades (Raffa et al., 2008). Regardless of how they operate, the frequency of natural disturbance events has been increasing around the world in recent decades as a consequence of climate change and anthropogenic alterations to both terrestrial and aquatic environments (Flannigan et al., 2000; Johnstone et al., 2016; Seidl et al., 2017). The rise in global temperatures, coupled with loss of biodiversity due to natural resource extraction and land use change, have heightened the impacts that disturbances exert on a number of ecological processes, sometimes causing ecosystems to transition into novel and often undesirable states (Parks et al., 2016). The continued sprawl of urban areas into natural environments has also amplified the number of people at risk to natural disturbances (Liu et al., 2015). As a result, environmental policies, particularly those aimed at minimizing disturbance impacts through mitigation or adaptation strategies, are being critiqued for their inability to ensure both long-term sustainability of natural resource use as well as address the risk of disturbances to human populations (Keskitalo et al., 2016; Six et al., 2014).

Researchers conducting environmental policy analysis can critically examine both discursive and substantive elements of disturbance-related policies as they seek to understand the degree to which these are informed by various scientific and political perspectives. Recent scholarship in environmental policy analysis suggests that ecological

* Corresponding author.
E-mail address: tcrnma3@ucl.ac.uk (M. Altaweel).

https://doi.org/10.1016/j.ecoinf.2019.02.014
Received 21 December 2018; Received in revised form 28 February 2019; Accepted 28 February 2019
Available online 04 March 2019

disturbances may be particularly fertile ground for opposition over what is termed "problem definition" (Fifer and Orr, 2003)—including the scope and urgency of the issue, causal factors, and culpability (including human actors, policies, or practices)—and the policy responses that logically flow from particular problem definitions (Abrams et al., 2018; Keskitalo et al., 2016; Morehouse and Sonnett, 2010; Prentice et al., 2018). Government agencies, non-government organizations, commercial interests, and other entities are likely to have strong motivations to attempt to shape public understandings and perceptions in the wake of large disturbance events (Boin et al., 2009; Keskitalo et al., 2016).

Of the various approaches to environmental policy analysis that currently exist, content analysis has emerged as a foundational approach, focusing attention on the discursive construction of popular understandings of complex ecological issues (Arts, 2012; de Jong et al., 2012, 2017; Kleinschmit et al., 2009; Leipold, 2014). To date, the majority of such analyses have taken a qualitative, deeply contextual approach facilitated by expert knowledge of key narratives as well as the deployment of and relationship between terms that are used in the construction of policy discourse. While such knowledge is necessary for focusing analysis on specific policy components, it is not without the risk of biasing the types and quantity of terms used in the analytical procedure. For instance, terms used in assessment often require prior knowledge of what those terms are and what topics are most relevant in policy-focused documents. Effectively, this could narrow the focus towards specific terms known to analysts rather than determining relevant terms based on document content. Furthermore, current methods typically involve a hybrid of manual and basic computational processes to scan documents in search of terms. Such approaches could be limited in the breadth or number of documents searched as well as in the depth of analysis undertaken by teams for long documents, potentially limiting their effectiveness for policy analysis.

To address these issues, this paper demonstrates the use of a semi-automated content analysis approach for investigating policies directed towards ecological disturbance. The goal is to demonstrate how our applied method has wider utility for analyzing policy text in relation to important environmental disturbances through the discovery of relevant terms and topics. We do so by treating policy-related documents as data with important ecological content concerning the framing of ecosystem dynamics in the context of ecological disturbances. We apply a methodology that conducts a quantitative analysis of texts contained in a large number of documents. Specifically, the methods presented here deploy latent Dirichlet allocation (LDA) and hierarchical Dirichlet process (HDP) topic models (Boyd-Graber et al., 2017) along with coherence modeling (Röder et al., 2015) in order to quantitatively analyze large volumes of text and to provide various metrics of association, proximity, and abundance of key terms in the documents. Topic models help determine relevant terms and topics covered by documents analyzed. Both LDA and HDP are commonly applied topic models that provide insight into corpora discussion (Yau et al., 2014*)*. Once topic models are used to determine relevant topics and terms, a focused term frequency–inverse document frequency analysis (Salton and Buckley, 1988) on documents allows us to investigate the context in which terms appear over time and the different document types in which they appear. The overall methods integrate machine learning, quantitative, and semi-automated approaches in order to evaluate environmental policy over time.

Applied content analysis in this paper is used on documents pertaining to mountain pine beetle (*Dendroctonus ponderosae* Hopkins, hereafter MPB), a bark beetle native to the pine forests of western North America. The most recent outbreak of MPB in North America represents one of the largest ecological disturbance events to forests on record, as it has led to the mortality of tens of millions of trees in Canada and the US (Rosenberger et al., 2017). MPB outbreaks have been recorded in previous decades throughout this area, but the most recent outbreak, which started in the mid-1990s and peaked in the early to mid-2000s, was unprecedented in both the level of tree mortality and in its spatial extent. Government agencies at varying levels have developed a range of policies in an attempt to mitigate MPB-induced impacts by, among other things, supporting regional response coalitions and expediting tree harvesting (Abrams et al., 2017; Davis and Reed, 2013). This paper attempts to understand how various government agencies in the US have framed and communicated issues related to MPB outbreaks. In the course of demonstrating the applied methods, we examine texts where MPB outbreaks have affected (explicit or implicit) policy discussion in government records. This complements other analyses of more limited scope that have examined the content and policy narratives associated with bark beetle-related legislation specifically (Abrams et al., 2018; Six et al., 2014).

2. Background

2.1. Content analysis

We define content analysis as an approach focused on inferences and interpretation of communication data using manual or computational methods (Krippendorff, 2013). Machine learning techniques have significantly influenced how approaches have addressed content analysis (Chau and Chen, 2008). Machine learning can be defined as the application of artificial intelligence (AI) techniques where computers can use what is learned in models or methods of inference, often through numerous iterations of sample data, in performing a given task (Michalski et al., 1983). Techniques in machine learning have enabled the discovery of relevant content patterns that includes the classification of texts either through categories created by users or categories that are determined from a corpus of texts (Gong and Xu, 2007). Techniques related to classification have tended to incorporate supervised and unsupervised methods, where supervised techniques require input from researchers and unsupervised techniques use minimal or no input for determining how texts can be classified. Topic modeling has emerged as a study area within machine learning that incorporates techniques where the goal is to determine what topics are being considered and how topics relate to one another in a corpus. Topics are defined as the focal semantic subjects that are relevant in documents and are often associated with multiple terms (Boyd-Graber et al., 2017).

2.2. Natural language processing in ecology

The application of natural language processing (NLP), which uses computation to understand natural language patterns and understanding within text, is frequently applied to content analysis. Techniques using NLP for environmental policy analysis have been steadily growing. Use of NLP by researchers has focused on, among other topics, how discourse is shaped around areas of ecological change, such as in natural resources, pollution, and climate change (Antrop, 2001; Pascoe et al., 2016). This has included evaluating public perceptions from sources such as newspapers (Altaweel and Bone, 2012) or social media (Veltri and Atanasova, 2017), using approaches that include thematic analysis and term relationships over time. In the area of management of resources, named entity recognition, which classifies and finds items or entities of potential interest in unstructured text, has been another method applied that helps to discover, that is find without user input, relevant actor relationships in discourse (Murphy et al., 2014). Here, actor relationships include people involved in discussion or making decisions relevant to environmental policy. Other research has focused on sentiment analysis that investigates perceptions or opinions of those who experience a natural setting (e.g., Becken et al., 2017). To date, topic modeling has not been extensively used in environment policy areas. Recently, Cheng et al. (2018) applied a topic modeling approach to assess the intersection between the ecology, environment, and poverty in order to improve understanding of sustainable development. The work demonstrates the potential for topic modeling applications in the area of policy.

One advantage topic models provide is the possibility to discover

relevant terms, where a limitation has been the lack of methods that determine, without a priori knowledge, relevant terms in studying environmental policy. Methods have generally used subject matter experts, where analyses might be limited by the capacity of the efforts and scope of coverage within corpora. However, in addition to assessment of content topics, there is also a need to understand how given topics or terms related to topics change over time. For instance, this was done in relation to land use issues, where terms related to land use change demonstrated activities affecting given regions (Altaweel et al., 2010). Combining techniques that discover relevant topics and terms, such as topic modeling, and assessments that track the relevance of terms for given topics over time can potentially allow analysts to better comprehend the relevance and change of given topics (Jelodar et al., 2018; O'Callaghan et al., 2015), including in the context of disturbance events.

3. Methods

3.1. Data

Before discussing the method deployed, a brief summary of the data is given. Data for this study consisted of official, publicly available US government documents, which are a particularly rich source for analyzing the discursive elements surrounding the management of forests and associated disturbances (Bone et al., 2016; Rayner et al., 2013). We used a multi-pronged approach to investigate relevant material, finding resources using academic search engines with a targeted review of specific databases (Table 1). Searches were conducted in early 2017 and included documents going back to the earliest records in each database. In all cases, search terms used were *mountain pine beetle* or *Dendroctonus ponderosae*. For each document retrieved, we recorded the document title, HTML or PDF web address, source name, publication date, material type (congressional document, federal agency document, White House document, Government Accountability Office (GAO) document, or legal news), broad government level, specific government level, and any relevant committee or subcommittee of the entity that produced the document. Following the retrieval and storage of relevant documents from the web, duplicates retrieved were checked for and removed.

Because the federal government owns the majority of land on which MPB occurs in the US, the US Congress, as well as federal agencies such as the US Forest Service (USFS), have a special interest and responsibility in anticipating and responding to MPB-related impacts. As the seat of executive power, the White House (i.e., Office of the President of the United States) is often an important source for public documents meant to frame public policy issues in ways favorable to the administration in place at the time (Vaughn and Cortner, 2005). Executive-level documents tend to be nontechnical documents crafted for broad public audiences, often justifying a particular course of action within the executive agencies or making the case for a preferred regulatory or legislative change (Vaughn and Cortner, 2005). Land management agencies themselves, such as the USFS, are also sources of information intended for both general public and more interested public (e.g., forest industry, conservation NGOs, recreationists) audiences; some of this information may include more technical forest management detail. Because the USFS is a federal government agency with leadership appointed by the President, its policy documents will often frame forest management issues using narratives that support the administration's agenda (Bone et al., 2016; McCarthy, 2005).

By contrast, congressional documents are likely to reflect more heterogeneous perspectives. This is because, as a deliberative body with representation of diverse states and districts, Congress contains a diverse assortment of interests regarding forest and public land issues. Congressional hearings, which often relate to legislation under consideration, oversight of federal agencies, or attention to issues identified as problems, normally include testimony from key witnesses that may likewise reflect a diversity of opinions and perspectives. Finally, the Government Accountability Office (GAO, formerly General Accounting Office), an independent, nonpartisan federal research entity, produces documents in response to official requests that represent syntheses of policy-relevant information on particular topics. Our content analysis includes all of the aforementioned categories of federal policy documents, along with legal documents that may reflect a wider range of voices engaging in legal and policy debates over issues related to forest management in response to MPB outbreaks.

Fig. 1 shows the entire corpus analyzed over different time intervals; overall, there were 1416 documents in the final database. While in more recent years there are more documents (a), the total number of words (b) peaked between 1975 and 1979 and pre-1960 words also indicate much longer documents.

3.2. Content analysis approach

The content analysis (Krippendorff, 2013) used here focuses on

Table 1
Databases and associated filters used to collect relevant policy documents from online sources.

Database name	Database link	Filters applied
Federal Digital System (FDsys) (U·S. Government Publishing Office)	https://www.gpo.gov/fdsys/search/home.action	None
Federal Register	https://www.federalregister.gov/	None
GovInfo: Beta Launch (U.S Government Publishing Office)	https://www.govinfo.gov/	Searched the “Congressional hearings”, “Congressional record”, “Congressional reports”, “GAO reports and comptroller general decisions”, “Journal of House of Representatives”, and “other” collections
Lexis Advance Research[a]	https://advance.lexis.com/	Jurisdiction: “All Federal”; Content type: “Statutes and legislation”, “Administrative materials”, “Administrative codes and regulations”, “Legal news”
Library of Congress	https://www.congress.gov/	Searched “all sources”
ProQuest Congressional	http://congressional.proquest.com	Narrowed search to: “Congressional record bound”, “CRS report”, “Hearings unpublished”, “House and Senate journals”, “House and Senate Reports”
Regulations.gov[b]	https://www.regulations.gov/	None
U.S. Department of the Interior: Office of Hearing and Appeals	https://www.doi.gov/oha/organization/ibla/Finding-IBLA-Decisions	Searched within “IBLA Decisions (1970-Present)”
U.S. Government Accountability Office	http://www.gao.gov/browse	None

[a] Some documents retrieved from Lexis Advance were “Notice” documents that simply provided the titles of other source documents (e.g., Environmental Impact Statements). In these cases, we conducted a Google search and then a separate search of an EIS document database via the Environmental Protection Agency's (EPA) website to find the complete document.

[b] In some instances, a chapter or subsection was pulled from a larger Environmental Impact Statement. When this happened, a Google search was conducted to find the complete document.

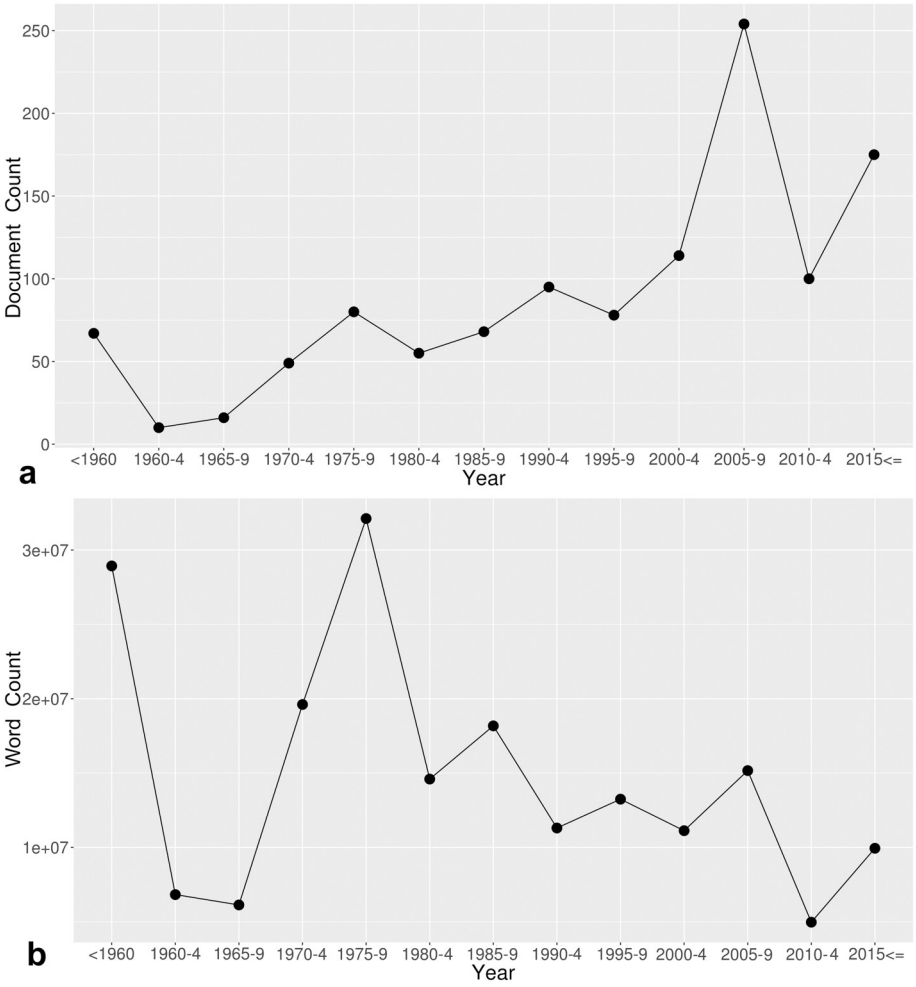

Fig. 1. Total documents analyzed (a) and their overall number of words (b) for different time intervals.

understanding how federal government entities have represented MPB issues. Several sets of analyses were conducted. The first applies topic modeling, which searches for collections of word associations that form topics within a text corpus (Blei, 2012). Topic modeling enables groupings of terms where each set of terms (e.g., law, land, area, forest) formulates a given topic (e.g., forest policy) evident in an analyzed set of documents. Documents can have multiple topics, where the analysis demonstrates the strength of given term associations for a given topic (Alghamdi and Alfalqi, 2015). Topic models effectively help classify texts based on what topics are discussed, often using machine learning techniques such as the methods applied in the approach discussed below.

3.2.1. Latent Dirichlet allocation

Topic modeling can be used to determine terms that have strong co-association and occurrence within topics. Effectively this means that groups of terms and their associations often help demonstrate what topics of discourse are of focus in texts. One type of topic model, deployed here, is LDA (Blei, 2012; Blei et al., 2010). LDA is an unsupervised machine learning technique that deploys a parametric approach that investigates groupings of terms and looks at the strength of associations between terms using generative probability. What LDA does is assume that documents consist of topics (e.g., insect management in ecology) with which terms are associated. Generally, words have a higher probability in belonging to a given topic when they co-occur frequently, where this pattern can be learned from texts used to train the algorithm. The clusters of terms and applying statistical association of words analyzed allows a determination that a given grouping of words addresses a given topic. For instance, the topic tree mortality could be more frequently associated with mountain pine beetle or fire; the fact that these terms appear together frequently shows that the topic tree mortality could be associated with MPB and fire. There can be multiple topics and the analysis determines how many topics are potentially evident; the subject expert may have to use their expertise in giving these topics a name such as tree mortality, as the topic determined from the analysis is generally abstract. When applying LDA, the analysis also requires a defined number of topics to search for. The number of topics can be changed in the analysis, allowing the strength of word associations to vary. This could suggest that a given number of topics might be weak or not accurately represent likely topics in the corpus. The process, therefore, is often done iteratively to find a more suitable number of topics.

In addition to the description of LDA above, we provide notation of the steps involved. First, the key variables in the applied model are:

α: scaling parameter of the Dirichlet prior document topic distributions,
d: document in corpus D.
β: parameter of the Dirichlet prior on the word distribution for a topic,
θ_m: topic distribution for a document m,
φ_k: the word distribution for topic k,
$z_{mn:}$: topic of the n-th word in document m,
w_{mn}: a word in document d.

The key input variable is w, while the others are latent variables,

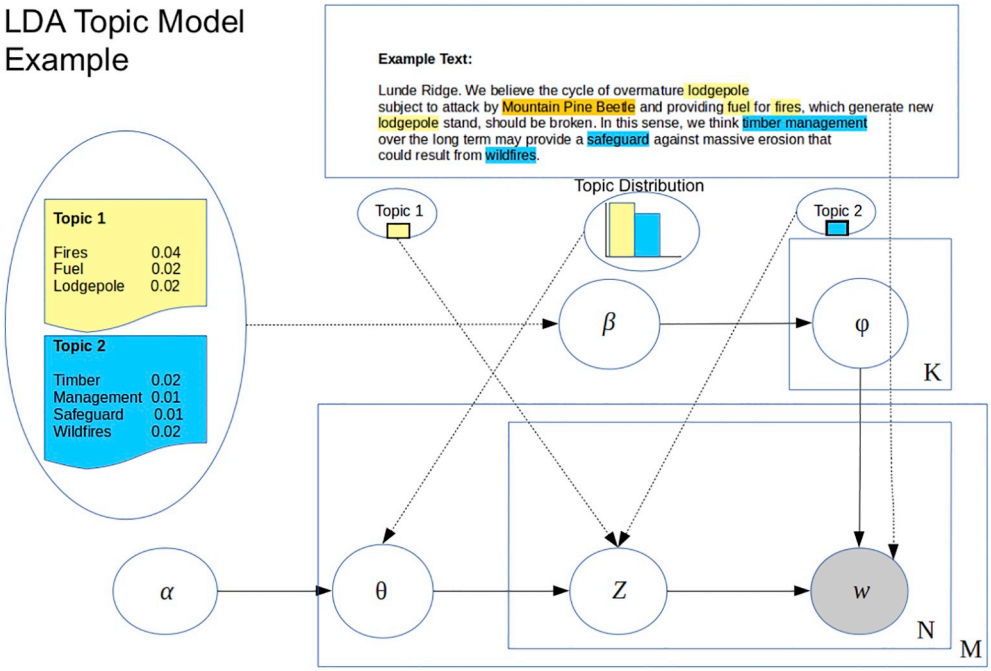

Fig. 2. Plate notation of the latent Dirichlet allocation process applied to an example text. The letters M, N, and K represent the number of documents, words, and topics respectively.

where a Dirichlet prior is used to model word distributions used to create the topic model. The generative process can be summarized as:

$$\begin{aligned} &\forall\ d \in D: \\ &\theta_m \sim Dir(\alpha) \\ &\varphi_k \sim Dir(\beta) \\ &z_{mn} \sim multi(\theta_d) \\ &w_{mn} \sim multi(\varphi_{zd}) \end{aligned} \tag{1}$$

where Dir() draws from a uniform Dirichlet distribution and multi is a multinomial. This can be summarized in Fig. 2, where the steps in the LDA are applied in the order given and on an example text. Both θ and φ represent matrices in the decomposed document representing the topic and word distributions respectively. The Dirichlet distribution is needed as a prior distribution for multiple terms and topics in a given document. Selection of the topic is then made from the multinomial distribution with a word drawn from the overall word distribution. The end result is an association of terms that relate to a given topic number. That topic number could be named by experts upon seeing that the word associations (e.g., timber, management, safeguard) appear related to, for instance, a forest management topic. What is powerful about this approach is that it determines word associations and frequencies in relation to topics without any prior input for terms to search. In other words, it generates the relevance of words in given topics and informs which words are associated in relation to a topic. It is also possible to visualize the topic model scores using a relevance rating for terms to topics. This allows one to visualize how close topics are to each other, that is similarity, and determine terms that relate to topics, where the intertopic distance is calculated using multidimensional scaling as discussed in Sievert and Shirley (*2014*). This effectively allows one to see how related topics are to each other and the terms in which relate to given topics.

3.2.2. Coherence modeling

The output of LDA provides term associations and strengths of these associations in determining a given topic. One problem with a topic model such as LDA is that a priori it is difficult to determine the number of topics to assess. For instance, one can estimate that there might be 100 topics, but potentially only a smaller number has any clear meaning or strong association between words. Coherence modeling can be applied to evaluate different topic models and determine topic strengths, which estimates an approximate number of likely topics. Coherence models can evaluate outputs from LDA, looking at scores within term associations, and produce an output that indicates if the number of topics is stronger or weaker. Effectively, a coherence model helps to determine the optimal number of topics for the given corpus based on how well topics score (Srinivasa-Desikan, *2018*).

Here we applied the approach by Röder et al. (*2015*), which is employed within Gensim (see below). The coherence model was used to determine the number of topics in the overall corpus and sub-analyses for different types of documents and years analyzed. The work pipeline in this approach includes first segmenting words from a corpus (*t*), that is, all words were placed in word pairs (S). Based on these segmented words, then a probability (*P*) was determined for given words based on a reference corpus, which can be text from the main analyzed corpus or other documents used to train the analysis. Then a confirmation measure (φ) was used to calculate the agreement between probabilities and word pairings. This effectively tried to determine the strength or agreement for word pairings. Finally, all the subsets for the given word pairings were aggregated to give a single coherence score (c). The workflow is summarized in Fig. 3.

3.2.3. Hierarchical Dirchlet process and topic model

A second machine learning topic model, HDP, was applied as a way to search for alternative term and topic distributions. Overlap between the methods provided greater confidence in the results, while allowing us to determine if other possible term groupings were evident. In this case, HDP is a nonparametric, unsupervised Bayesian process, which is an application of a randomized Dirchlet process; this means it uses prior knowledge of probability distributions for clustered groups of data comprised of terms and topics. The distributions were then used to create term and topic associations similar to LDA (Teh et al., *2006*). The main difference with LDA is that the number of topics is not required as an input, where the output could be used to determine a likelihood for the number of topics. However, as both LDA and HDP apply probability sampling, one approach was to integrate them together to strengthen overall confidence in terms captured for given topics.

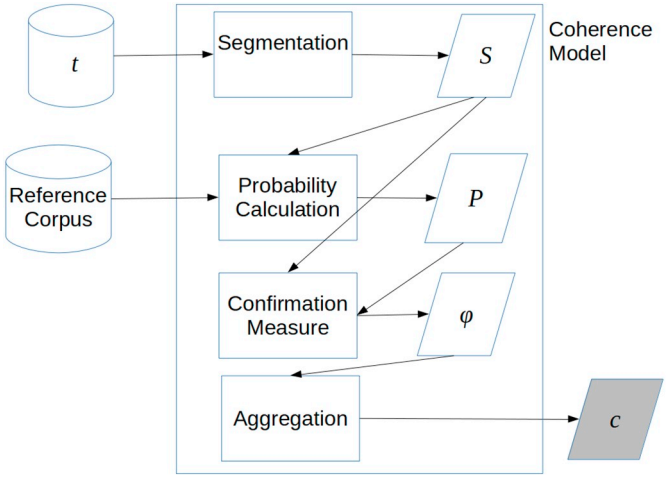

Fig. 3. Workflow for the coherence modeling.

First, we summarize the main process for HDP. The method uses the general application:

$$G_0 \vee \alpha, H \sim DP(\alpha, H) \tag{2}$$

where a Dirichlet process (*DP*) applies α, a concentration parameter, on the base distribution (G_0) that varies around H, a base probability measure that provides the prior distribution. To determine the the jth grouping of terms, the Dirichlet process applied the following distribution:

$$G_j \vee \alpha_0 G_0 \sim DP(\alpha_0 G_0) \tag{3}$$

where the G_j distribution for groupings is determined by a Dirichlet process where it is governed by α_0, a concentration parameter. This then enables a hierarchical Dirichlet process to be created that can use a prior distribution on the actors for given grouped data based on the following:

$$\begin{aligned} \theta_{ji} \vee G_j &\sim G_j \\ x_{ji} \vee \theta_{ji} &\sim F(\theta_{ji}) \end{aligned} \tag{4}$$

where θ_{ji} are parameters that are determined from the G_j prior distribution; each individual observation or term (x_{ji}) is given through a distribution *F* with an associated parameter (θ_{ji}). This provides a hierarchy model of associated topics and terms. Topics are inferred, resulting in the total topic numbers being determined as part of the results. Fig. 4 summarizes HDP for each data item. One potential approach is to combine LDA and HDP, as both are probabilistic models where each model alone may not fully capture potential term associations. In this case, where topics are determined, terms associated with topics could be combined from the two outputs so that the overall terms present broader coverage for a given topic. *Overlap* between terms demonstrates that the topic is the same or similar, while additional terms help to show other potential terms relevant for the topic.

3.2.4. Term frequency–inverse document frequency

Topic models help determine what topics are relevant in discussion. However, this assessment is done for the entire corpus analyzed. By themselves or without modification, topic models do not look at changing term usage over time or in segments within a given corpus. For instance, outputs may show the topic of tree mortality is relevant, but terms for that topic may not be consistently used or discussed in relation to that topic. This could reflect changes in semantics or even change in the relevance of the topic over time.

Based on this, and in addition to topic modeling, term frequency–inverse document frequency (tf-idf; Salton and Buckley, 1988) analysis was used to determine relevance of terms as they changed over time. Although tf-idf is now a relatively old method, it is still powerful

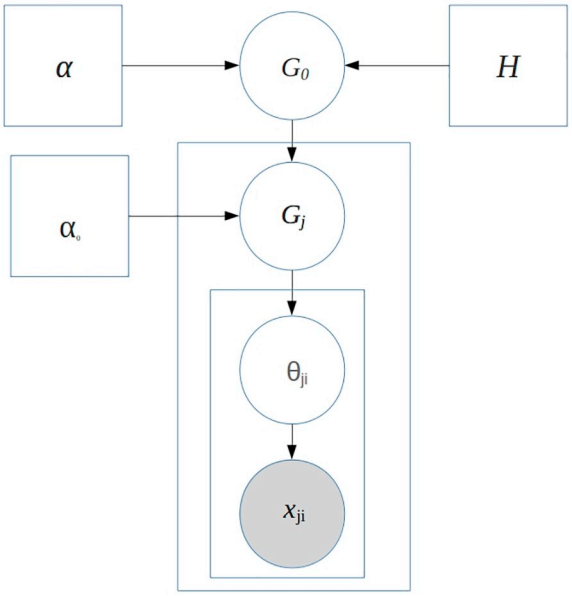

Fig. 4. The workflow for the hierarchical Dirichlet process applied.

as its basic approach shows the relevance of terms in documents and can be used to analyze this relevance in different parts of the corpus (e.g., see Walter et al., *2017*). The method can also be used along with topic modeling, where terms discovered utilizing this approach could then be searched for their change. Effectively, topic modeling allows us to discover terms associated with topics; tf-idf could then be used to monitor the terms' relevance across the temporal range of the corpus. Topic modeling is useful for indicating important topics and associated terms, but tf-idf allows us to more directly focus on the terms themselves, which may intersect multiple topics. For instance, the relevance of fire could be determined by a topic model, but tf-idf tracks how it is used over time, helping to find when the term is more frequently used. The basic notation for tf-idf is as follows:

$$S_{td} = TF_{td} * IDF_t \tag{5}$$

where *S* is defined as the tf-idf score of a term (*t*) within document (*d*); *TF* is the term's frequency (3); *IDF* is the inverse document frequency (4). *TF* is determined by:

$$TF_{td} = \frac{n_t}{\sum w_i} \tag{6}$$

where the number of instances (*n*) for a term (*t*) is determined. Leading to the final part of the calculation, that is IDF:

$$IDF_t = log\left(\frac{N}{1 + |(d: t \in d)|}\right) \tag{7}$$

where *N* is the total documents assessed and the denominator represents the number of documents (*d*) that have *t*.

3.3. Content analysis steps

The first step undertaken was to investigate the entire corpus of documents (PDFs) that were downloaded and relevant to MPB. This entailed topic modeling using LDA, where the optimal number of topics was determined by using the coherence model and looking at the topics covered. A loess regression was applied to show the general trends in results where the number of topic models and their coherence scores vary from one input topic number to the other.

After this step, we decided to focus more specifically in parts of text that mentioned MPB using the terms mountain pine beetle(s) or the scientific name *Dendroctonus ponderosae*. In this case, each sentence that included such terms in a document, including the sentence before and

sentence after, were kept for the analysis. This resulted in a different set of topic model outputs to which we also applied coherence models to determine the best number of topics to analyze. Similar to before, LDA was applied along with a coherence model for determining the number of topics. Additionally, HDP was used to determine relevant topics and terms, where results could be combined with LDA. The intent with this step was to determine if a more focused search would yield more relevant topics in relation to MPB.

In the next step, texts were divided in the following time increments: 1960 < , 1960–1964, 1965–1969, 1970–1974, 1975–1979, 1980–1984, 1985–1989, 1990–1994, 1995–1999, 2000–2004, 2005–2009, 2010–2014, and 2015 ≤. This is done because not every year has the same number of documents; aggregating in some way helped to look at term trends while discounting year-to-year document number variations. A tf-idf analysis was then applied to determine how terms from the LDA/HDP topic model outputs differed over time. Furthermore, texts were broken down into categories based on different government branches and subdivisions based on the types of government texts discussed earlier.

3.4. Applied tools

Tools used for this analysis included the Gensim and NLTK (Rehurek and Sojka, *2010*) libraries executed in Python (2.7+). The Gensim toolkit contains LDA, HDP, and coherence modeling used here; NLTK was applied to prepare documents and remove stop words. Gensim was also used to lemmatize analyzed words; this means grouping together different variant forms for words and using them all in the analysis rather than the single term. Words were parsed using NLTK built-in method that breaks up sentences; Gensim is used to group variants of terms (e.g., disease = disease, diseases, diseased). Another key package used is the LDAvis package (Sievert and Shirley, *2014*), which was used for visualization of topic models. The applied code and most of the data are provided in GitHub (*2019*) and as a data link in this paper (see Dataset). This also provides the metadata that discusses the texts analyzed, which could be downloaded. Some of the analyzed PDFs are provided, but the entire corpus is too large to include.

4. Results

4.1. Topic modeling of all documents

The first step was to assess the entire corpus to see what topics were focused on and terms that appear to relate to key topics. Fig. 5a depicts the coherence modeling in determining the number of topics covered with a loess regression on the results. Overall, this shows wide disparity based on the number of topics tested; however, it generally shows that the coherence score increases as the number of topics increases, suggesting it is more likely that around 140 topics were covered. The highest score was at 27 topics (score = 0.37), but overall trends suggest a greater number of topics were more likely based on the fact that greater topic numbers produced higher scores. Fig. 6 shows LDA intertopic distances or similarities for 70 topics. Terms such as *forest*, *service*, *state*, *land*, *area*, *act*, and *fund* were the most common terms overall. This reflects a large focus on areas, such as forests, but also legislation, including their funding. The overall distribution of the distances between topics, and high number of topics, showed a wide variety of coverage and divergences in term composition among topics.

The topic results demonstrate that there were likely to be many topics covered by the documents analyzed. While we could utilize these results, they also inform that one could capture many topics well outside of MPB given the results. Similar to other approaches (Polatkan and Nieselt, 2013), we have focused our search to literature most likely related to our interest area to see what topics were more likely related to MPB. Limiting the focus on the more likely relevant areas potentially narrows the topic coverage, providing a greater focus in relation to MPB. This was accomplished through a keyword approach that finds parts of text that are more likely to be relevant. This provides more descriptive and relevant parts of texts which could then be further analyzed for their topics using the automated topic modeling approach described (Ahonen et al., 1998). As stated previously, the corpus was analyzed and limited to only sentences where MPB is evident, using *mountain pine beetle* and *Dendroctonus ponderosae* as terms, with sentences prior to and after a given MPB sentence kept for analysis, with the texts then assessed for their coherence scores using LDA. Fig. 5b shows the result of this, showing that topics were more coherent or have a higher coherence score when far fewer topics are assessed (at about 10 topics; Fig. 6). In other words, coherence modeling on a more limited corpus showed far fewer topics than assessing the overall corpus, where 10 topics scored a relatively high score (0.37). The decline in the coherence score was consistent, whereas the entire corpus showed a general trend but also showed greater variability in consistency of an increasing coherence score. The more limited and focused MPB-based search was more reliable in its coherence scoring, indicating that texts were more focused in topic coverage when MPB was searched.

Applying the same visualization on the relevance of topics and their associated terms on a more limited 10 topics demonstrated several notable results (Fig. 7a). First, topics 2, 3, 5, and 8 are the most similar to each other, showing that within the 10 topics there were similarity and overlap in coverage. Topic 1 had the largest composition of the top terms (Fig. 7b), where the tokens, that is the terms, were at 34%. Not only do we see *mountain*, *pine*, and *beetle* appearing frequently, as to be expected, but terms such as *bark*, *area*, *outbreak*, *fire*, *insect*, and *control* were among the most common. These terms appeared across multiple topics. To further capture relevant terms and strengthen the LDA approach, HDP was applied. As discussed previously, the LDA and HDP terms, for the first ten topics, were combined. Table 2 reflects the integration of LDA and HDP terms found for the top ten topics, which had the highest coherence score, and the fifteen highest scored terms found using LDA and HDP methods. Terms directly related to MPB (i.e., *mountain*, *pine*, *beetle*) were removed from the table, as those terms were expected to be high in number. The ten topics are given names that relate to their terms in Table 2. This was done by having the authors jointly agree on the relevant titles of the terms. While this is subjective, it enabled an understandable topic reference to reflect the subject areas covered by the combined terms.

4.2. Topic modeling of sub-documents

The next step in the analysis was taking the terms from the topics and looking at the tf-idf trends. As stated, we sought to understand temporal change of identified terms within topics, determining which terms have likely become more important recently or were of greater use in earlier periods. Rather than simply taking the top terms, which may have been clustered around a few topics, top terms associated with the ten most common topics were selected. In this case, we chose terms related to key events, actions, or outcomes that relate to concerns on a given MPB topic. Finding relevant terms that related to given behaviors or outcomes and looking at them across time indicates where focus had been among the government-related literature (Popescu et al., 2011). Fig. 8 shows results of tf-idf analysis on the following terms that were among the most common and reflective of events, actions, and outcomes: *disease*, *fire*, *infestation*, *mortality*, *outbreak*, *attack*, *treatment*, *control*, and *management*. These terms potentially intersect multiple categories but reflect relevant interests in ecosystem disturbances and ways to prevent or mitigate these disturbances. The terms were chosen because linkages of outcomes, actions, and ways to address MPB reflected ecological disturbances that link to policy; these linkages reflect key concerns for policy that could aid in better focusing where resources are needed by decision-makers (Polasky et al., 2011).

Terms such as *disease* and *infestation* demonstrated less prominence

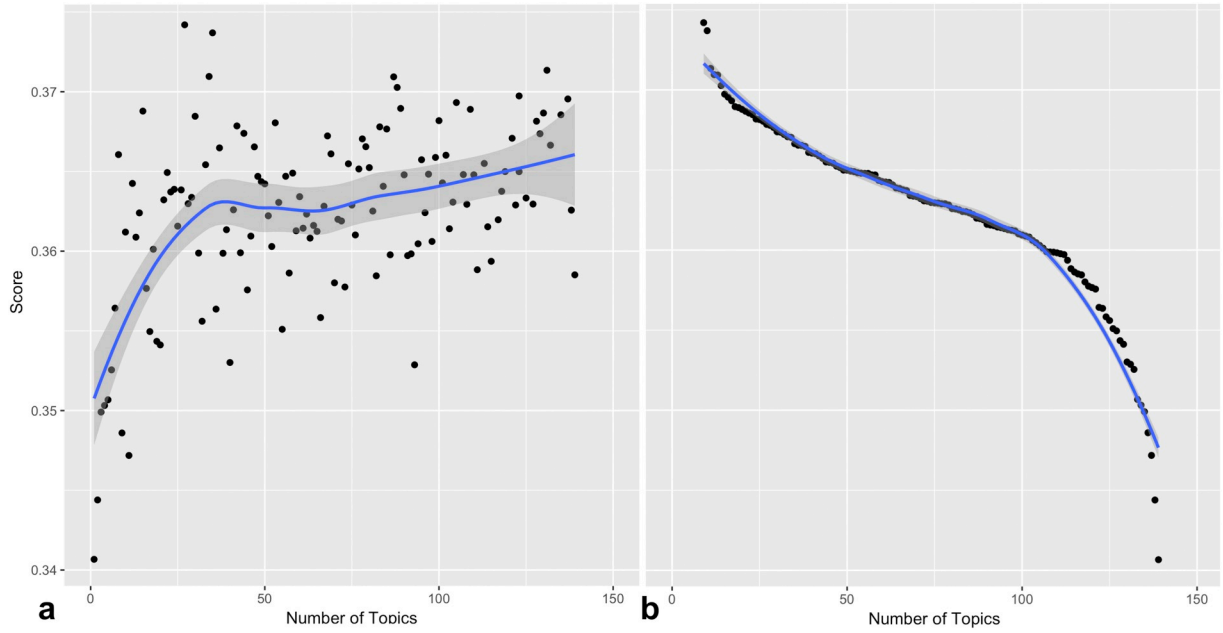

Fig. 5. Coherence modeling scores (a) for the entire corpus and the more limited (b) search on mountain pine beetle sentences, including sentences before and after mountain pine beetle terms was mentioned.

in texts over time, while *fire* generally increased, although fluctuations were evident. In fact, *fire* had generally become among the most common topic-based term for documents after 2000. *Mortality* and *treatment* also became more of a focus in recent time intervals, whereas they were almost never discussed in documents prior to 1960. The term *outbreak* was relevant in documents prior to 1970, but the term declined in relevance in documents, while then increasing in tf-idf score after 2000. Interestingly, *disease* had declined the most in relevance across time.

4.3. Federal agencies and congressional documents

Another way documents were assessed was by dividing them into relevant categories, based on the type of government documents

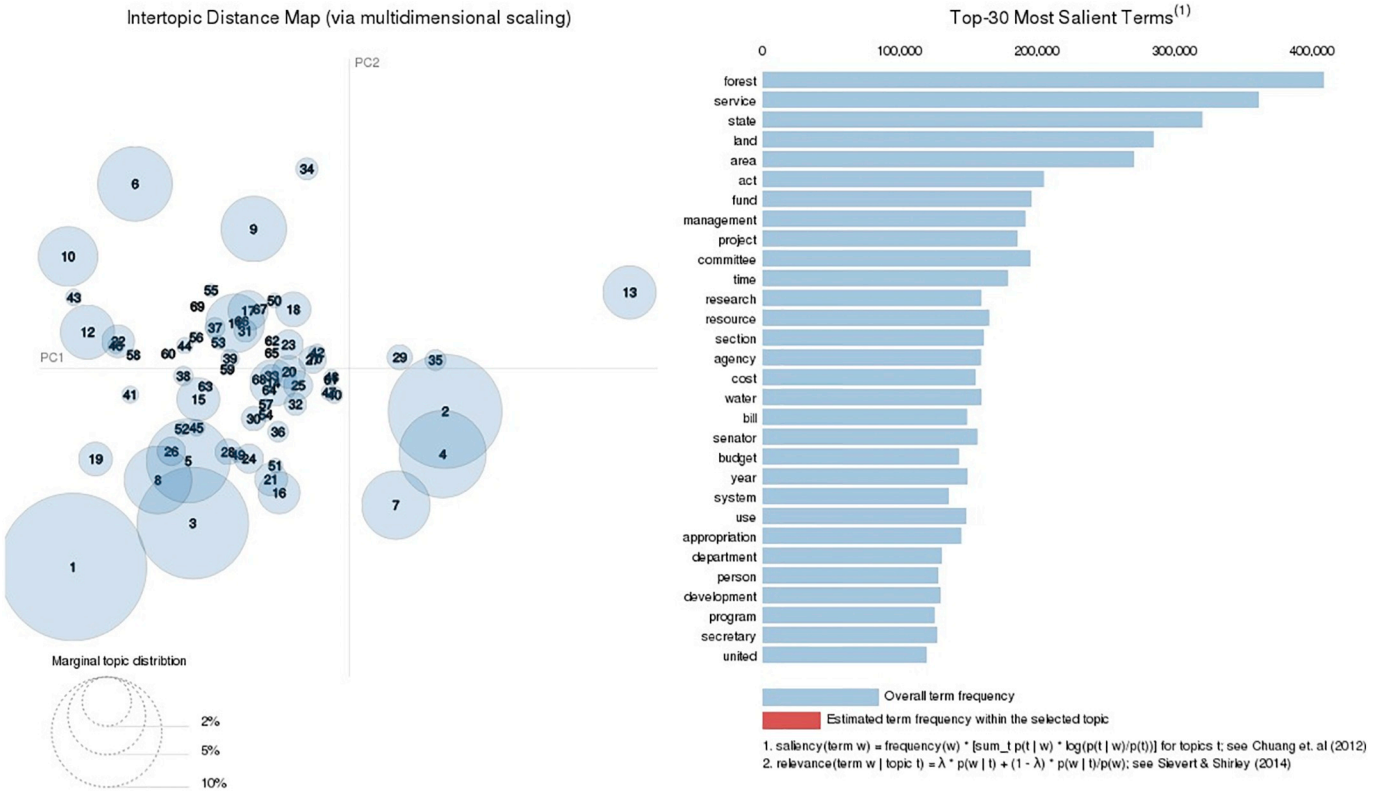

Fig. 6. Intertopic distances for 70 topics from the latent Dirichlet allocation results on the entire corpus. The results show the relationship of topics to each other (by distance) and variety of topics.

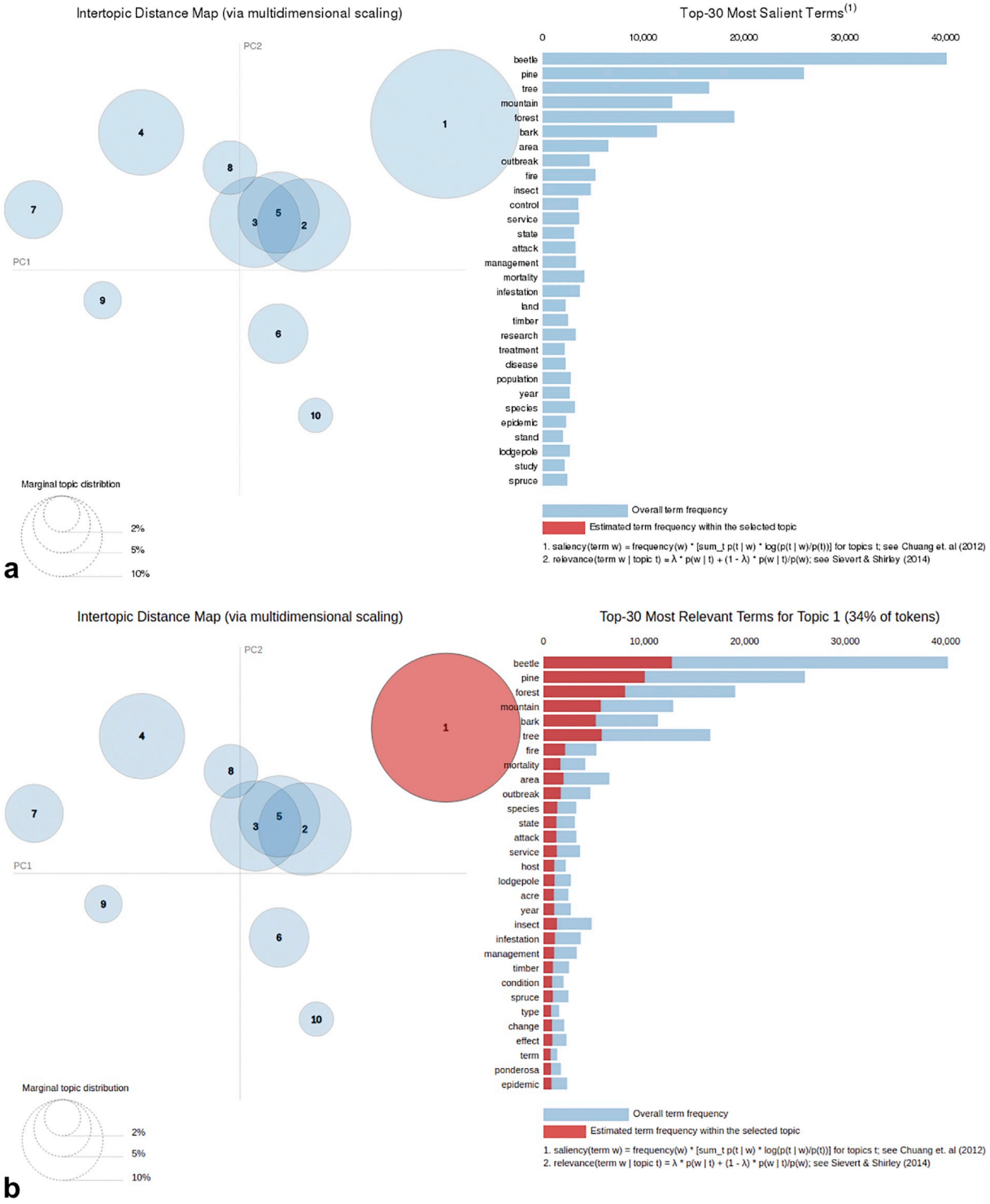

Fig. 7. Intertopic distances (a) for 10 topics from latent Dirichlet allocation results on the more limited mountain pine beetle search and highlighted (b) common terms from Topic 1, which had the most top terms in results.

available from the public data. Two of the major categories among all documents were federal agencies (e.g., Department of Interior; 650 documents out of the 1416 documents) and congressional documents (651 documents), which together made up the vast majority of documents. Other categories of documents, such as Government Accountability Office reports, Legal News, and White House documents, represented far smaller categories that could not be adequately assessed without significant gaps in the record over time. In other words, there were periods of significant temporal gaps for these three categories.

Fig. 9 shows the tf-idf scores for federal agency (a-c) and congressional (d-f) documents on the same terms as Fig. 8.

For federal agency documents (Fig. 9a-c), results indicated a large increase in focus on *fire* since 1960, although the focus peaked around 2010. *Disease* and *infestation* increased since 1960, but they peaked around 2000 and 1995 respectively. In fact, all terms analyzed increased after 1960, although the peaks for the tf-idf scores varied. For congressional documents (d-f), some general trends appear similar but the tf-idf values were different. One notable result was the term *control*,

Table 2
Top ten topics, including their suggested focus and titles as determined by the authors, and common associated terms based on *latent Dirichlet allocation* and *hierarchical Dirichlet process* analyses.

Topic	Topic name	Top topic terms
1	outbreak area	forest, tree, area, fire, insect, service, control, outbreak, management, research, population, attack, mortality, infestation, species
2	tree mortality	tree, forest, mortality, area, attack, outbreak, stand, insect, fire, plot, state, disease, infestation, service, species
3	research and services	forest, tree, area, research, state, mortality, control, management, fire, service, agency, acre, time, timber, forest
4	management	forest, tree, outbreak, insect, area, mortality, species, fire, management, host, term, search, states, climate, service
5	infestation	tree, forest, area, control, infestation, fire, outbreak, insect, attack, service, state, policy, land, management, condition
6	outbreak control	tree, insect, forest, area, outbreak, mortality, research, population, control, resource, pest, infestation, habitat, loss, epidemic
7	fire	forest, tree, fire, area, population, outbreak, insect, management, state, service, population, research, treatment, host, attack
8	insect control	forest, tree, insect, control, outbreak, fire, research, infestation, mortality, area, log, emergence, attack, treatment, temperature
9	outbreak factors	forest, tree, insect, area, outbreak, spruce, fire, control, infestation, disease, temperature, precipitation, variable, distribution, area
10	tree population	tree, insect, forest, area, outbreak, mortality, research, population, control, fire, rate, tree, effect, response, climate

which seemed to be a key focus from 1975 to 1979, before then declining; this was notable because this term was not as prominent for federal agency documents. *Outbreak* was a term that showed overall increase, with a peak between 2010 and 2015; *mortality* peaked at 1990–1994.

Another way to analyze these results, and determine key differences, was to look at term relationships across time. Looking at correlations in terms assessed, Tables 3-5 shows Pearson product correlations coefficients (r) for federal agency, congressional, and federal agency and congressional terms respectively. Overall, the only r values over 0.7 between federal agency and congressional terms pertained to the terms *fire* and *control*. On the other hand, many more strong correlations were noted when investigating only congressional or only federal agency terms. For instance, federal agency documents showed a relatively strong correlation ($r = 0.91$) for *management* and *disease*. Similarly, *disease* and *infestation* were strongly correlated in congressional documents. In effect, results showed that congressional and federal agency documents often did not correspond closely in topic coverage at a given time interval assessed; however, a number of terms within document-type (i.e., federal agency or congressional documents) did.

Additionally, we chose to look at *temperature* as a keyword, as it served as term that sometimes appeared in relation to text with MPB (Fig. 10). Other terms, such as *climate*, or even *climate change*, were far less frequently used. The term *temperature* served as an indication of how often weather-related terms came up in discussions where MPB was a focus. Fig. 10a shows the overall trend of *temperature* for all documents, while (b) and (c) reflect federal agencies and congressional documents, respectively. Overall, there was an increase in treatment of *temperature* in more recent periods, particularly after 2000. In 1990–1994, interestingly, there was a spike in discussion on *temperature*, after which it declined and then increased again from 2000 but declined in documents from 2015. For congressional documents, only after 2000 was there a larger increase in tf-idf scores on *temperature*. Overall, congressional documents had the highest score in relation to *temperature*.

5. Discussion

5.1. Benefits of approach

Policy analysis related to ecological disturbance is a complex undertaking. While searching for terms in policy documents is a relatively trivial process, providing context to such terms and understanding how this context changes over time present various theoretical questions (i.e., how to define context?) and technical challenges (i.e., how to measure and track context dynamics?). Conventional content analysis approaches steeped in qualitative reasoning of key terms and their appearance throughout documents have the potential to provide rich insights into ecologically related policies, but are subject to a narrow scope of analysis due to the need to provide key terms to search a priori, and because qualitative approaches are limited in terms of being able to connect multiple key terms to each other and over time.

Our study provides a step in overcoming these limitations by combining topic modeling with tf-idf analysis, where it has shown potential in understanding discourse in relation to policy analysis. In particular, topic modeling, such as using latent Dirichlet allocation (LDA) and hierarchical Dirichlet process (HDP), helped identify key topics and associated terms that appeared frequently and together in wider

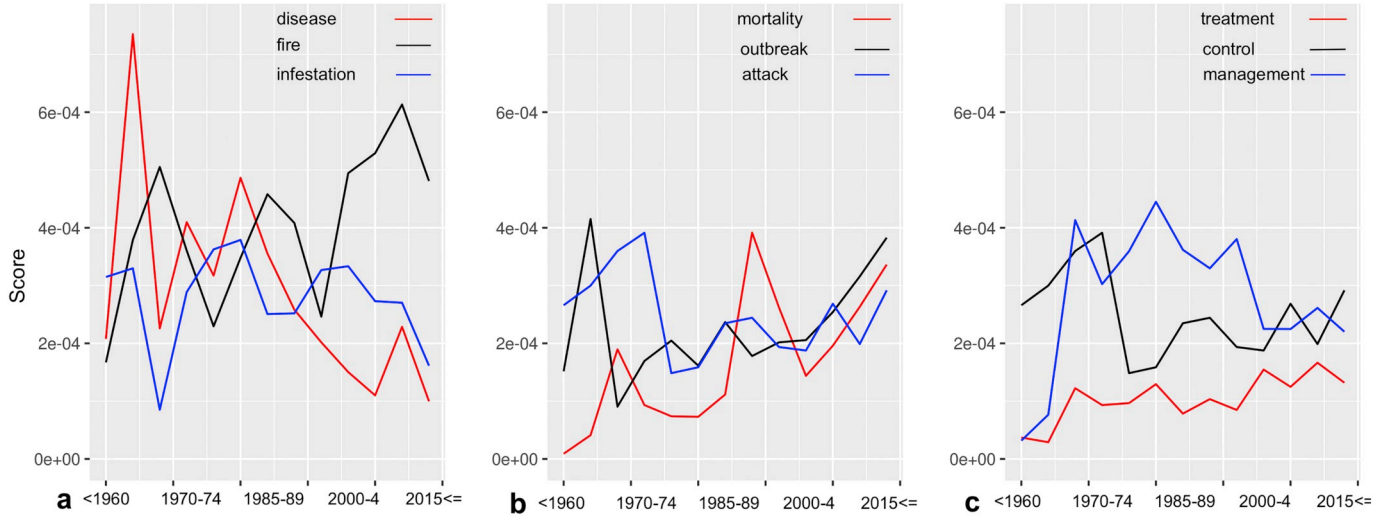

Fig. 8. Selected term frequency-inverse document frequency scores over time.

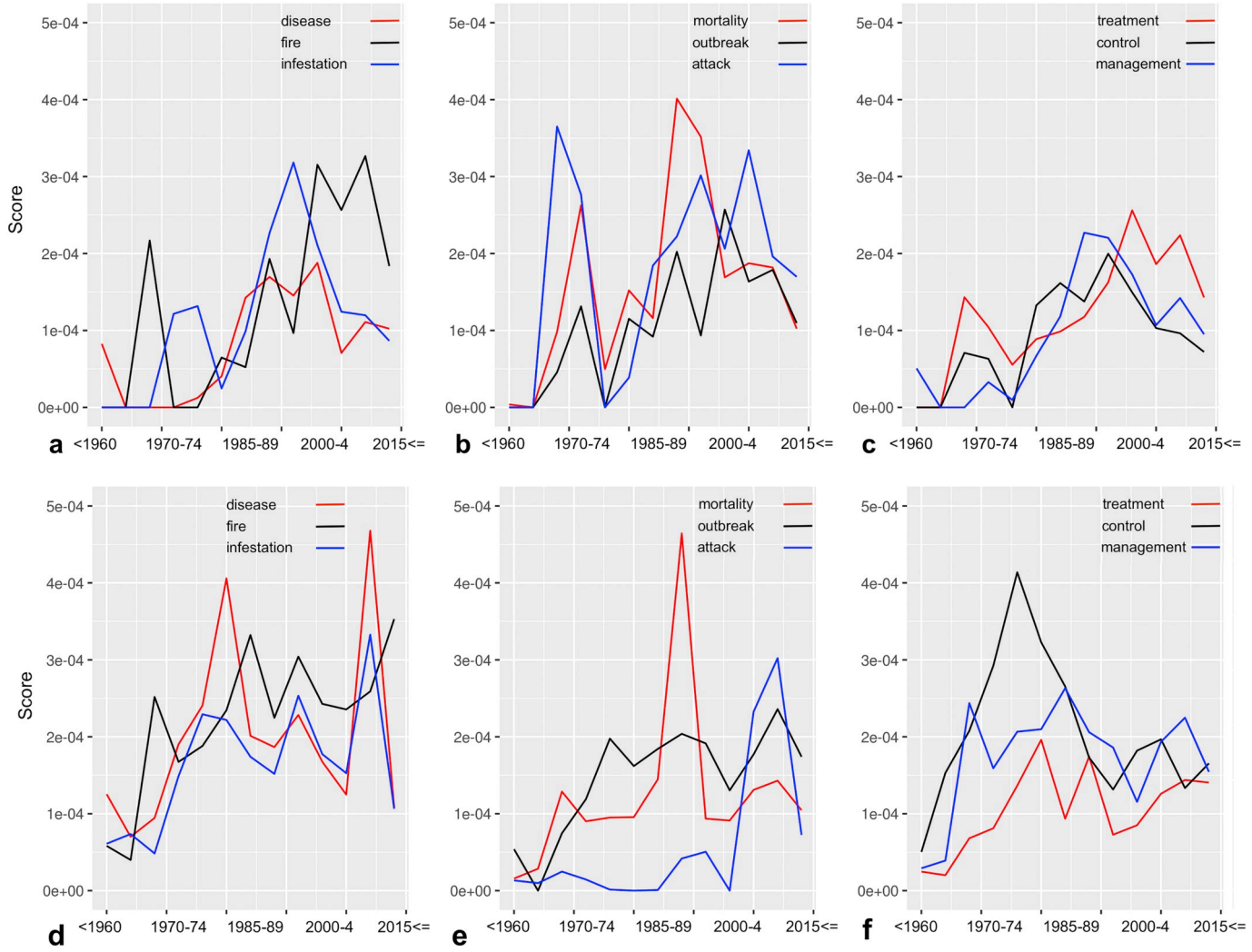

Fig. 9. Federal agency (a-c) and congressional (d-f) term frequency-inverse document frequency scores for selected terms over time.

discussions within a set of policy documents. This informed us as to what terms were worth investigating further without prior knowledge. These discovered terms also show potential topics or themes of discourse, demonstrating where the US government has been focused in relation to MPB. The analysis could then be carried further by investigating some notable terms from the topic modeling results. This then included using tf-idf to investigate the significance of specific terms across time, and enabling statistical relationships for terms with other terms in given document types, including categorizing documents further into groupings such as congressional and federal agency documents to see where variations in discourse are evident.

The results from this study demonstrate that mountain pine beetle (MPB) is discussed in relationship to multiple key terms when examining the entire corpus of government-related documents. This is expected as MPB, like many ecological disturbances, appears in the ecological discourse related to forest health (e.g., succession), biophysical processes (e.g., watershed hydrology), and other natural disturbances. Similarly, MPB is often used in the social discourse surrounding public safety (MPB-infested trees are considered hazards when falling after mortality) and social values (e.g., whether or not to cut trees to mitigate further MPB outbreaks).

While these findings provide useful insights, we were able to elicit more specific findings by narrowing the analysis to only sentences that incorporated MPB, including sentences nearby, which greatly reduced

Table 3
Federal agency-related document terms and their correlation using Pearson's r.

	Disease	Fire	Infestation	Mortality	Outbreak	Attack	Treatment	Control	Management
Disease	1.00	0.50	0.67	0.49	0.65	0.22	0.53	0.70	0.91
Fire	0.50	1.00	0.30	0.31	0.75	0.56	0.89	0.42	0.50
Infestation	0.67	0.30	1.00	0.80	0.56	0.43	0.55	0.67	0.83
Mortality	0.49	0.31	0.80	1.00	0.65	0.61	0.49	0.71	0.77
Outbreak	0.65	0.75	0.56	0.65	1.00	0.50	0.82	0.66	0.71
Attack	0.22	0.56	0.43	0.61	0.50	1.00	0.69	0.55	0.38
Treatment	0.53	0.89	0.55	0.49	0.82	0.69	1.00	0.64	0.59
Control	0.70	0.42	0.67	0.71	0.66	0.55	0.64	1.00	0.80
Management	0.91	0.50	0.83	0.77	0.71	0.38	0.59	0.80	1.00

Table 4
Congressional-related document terms and their correlation using Pearson's r.

	Disease	Fire	Infestation	Mortality	Outbreak	Attack	Treatment	Control	Management
Disease	1.00	0.25	0.87	0.10	0.62	0.31	0.60	0.25	0.45
Fire	0.25	1.00	0.40	0.30	0.72	0.49	0.52	0.16	0.74
Infestation	0.87	0.40	1.00	0.12	0.78	0.24	0.53	0.25	0.48
Mortality	0.10	0.30	0.12	1.00	0.49	0.44	0.56	0.02	0.46
Outbreak	0.62	0.72	0.78	0.49	1.00	0.52	0.77	0.28	0.71
Attack	0.31	0.49	0.24	0.44	0.52	1.00	0.60	0.52	0.59
Treatment	0.60	0.52	0.53	0.56	0.77	0.60	1.00	0.45	0.62
Control	0.25	0.16	0.25	0.02	0.28	0.52	0.45	1.00	0.50
Management	0.45	0.74	0.48	0.46	0.71	0.59	0.62	0.50	1.00

the topic coverage of documents. From these, relevant terms to policy-related areas, including *fire*, *outbreak*, *disease*, *infestation*, *management*, and *control*, among others, emerged as important terms that often occurred between different topics. The connection to these terms demonstrates the focus of government policy documents on MPB mitigation through methods of management and control, potentially in order to reduce risk to secondary disturbances such as wildfire. These findings are echoed in the topic modeling and coherence scores, which suggested ten topics as a reasonable number of topics covered. These topics are related to: outbreak area, tree mortality, research and services, management, infestation, outbreak control, fire, insect control, outbreak factors, and tree populations.

Using the topic modeling and coherence scores results, we were able to determine an increase over time in focus on terms related to *fire*, *mortality,* and *treatment*. Such findings are important because they demonstrate how our methods are able to track the frequency of terms over time as they relate to what is being discussed in government documents. Nelson et al. (2016) report that research on the relationship between fire and MPB outbreaks significantly increased starting at the turn of the century, mostly due to an increase in government funding to examine this relationship, and there has been increased concern that MPB attacked forests would be far more susceptible to catastrophic wildfires. Consequentially, the concern over increasing fire frequency following MPB outbreaks has led to a significant increase in forest treatments to mitigate MPB spread. The USFS has, in fact, proposed and implemented a number of projects designed to reduce forest susceptibility to uncharacteristic disturbance events—including both insects and wildfire (Bobzein and Alstyne, 2014; Six et al., 2014). The widespread promotion and use of such treatments is relatively recent, as many forests have demonstrated departures from historic disturbance patterns since the late twentieth century. Furthermore, while *climate change* had not been commonly applied in associations with MPB in our results, *temperature* had become more prominent in the MPB-related literature. The findings on *temperature* are notable as warmer temperatures in recent decades are directly attributed to increased incidence and severity of MPB outbreaks (Raffa et al., 2008).

Other noteworthy results show that federal agency and congressional discourse, or at least the use of similar terms by these government entities, often did not correspond closely in time (*fire* and *control* being an exception), suggesting that discourse between different branches of the government often had different interests. Terms such as *infestation* and *disease* were often discussed together in congressional documents, as an example, but they were not clearly discussed across some branches of the government. This could suggest different priorities by different parts of the government, or simply that different terminology is employed among agencies, which in itself would call for a reconsideration of the shaping of ecological disturbance discourse in government. Overall, the US government showed more interest in areas of more immediate concern, such as *fire and disease*, in contrast to long-term problems in MPB outbreaks. It was also concerned with management of resources as they were affected by outbreaks, which was expected, in particular where funding would be a key issue.

5.2. Limitations

Data limitations of the work included the scope being focused on by government-related documents. Ideally, work that incorporates the academic literature as well as government literature would be a better way to demonstrate linkages between MPB outbreaks, research, and policy. We attempted to procure research publications but paywalls by journals limited this possibility. As for the methods, the main limitation was the topic modeling deployed still required subjective interpretation, mainly with the term associations and topic designations. Additionally, LDA and HDP apply what is called a bag-of-words approach, which means sentence structure is largely ignored and words were treated as one grouping that disregards word order or grammatical elements (Zhang et al., 2010). Potentially, a method investigating sentence structure could be useful to determine cases where false positives for topics may skew the strength of topic and term associations. Furthermore, topic models were generally static, requiring significant modification to make them more easily usable for finding topical differences within various segments of the corpus. While we attempted to address this, in part, through tf-idf and by narrowing the search to MPB sentences, more dynamic interpretation could be beneficial in cases that can better analyze different segments of text over different time or relevant intervals. Despite these limitations, overall the terms found proved to be relevant to MPB and the fact that they emerged from the text rather than depended on expert knowledge made the overall

Table 5
Comparing federal agency (top row) and congressional (left vertical) terms' r values.

	Disease	Fire	Infestation	Mortality	Outbreak	Attack	Treatment	Control	Management
Disease	0.13	0.17	0.16	0.25	0.29	−0.13	0.28	0.31	0.25
Fire	0.49	0.49	0.45	0.42	0.46	0.55	0.66	0.72	0.51
Infestation	0.33	0.26	0.54	0.41	0.40	0.03	0.47	0.46	0.47
Mortality	0.44	0.34	0.42	0.68	0.50	0.32	0.23	0.39	0.57
Outbreak	0.49	0.40	0.61	0.56	0.51	0.28	0.55	0.56	0.61
Attack	0.11	−0.05	0.06	0.30	0.33	0.12	0.08	0.30	0.12
Treatment	0.20	0.33	0.23	0.41	0.48	0.05	0.34	0.36	0.33
Control	−0.41	−0.32	−0.04	−0.05	−0.10	−0.14	−0.11	−0.06	−0.36
Management	0.11	0.31	0.25	0.41	0.28	0.50	0.43	0.53	0.23

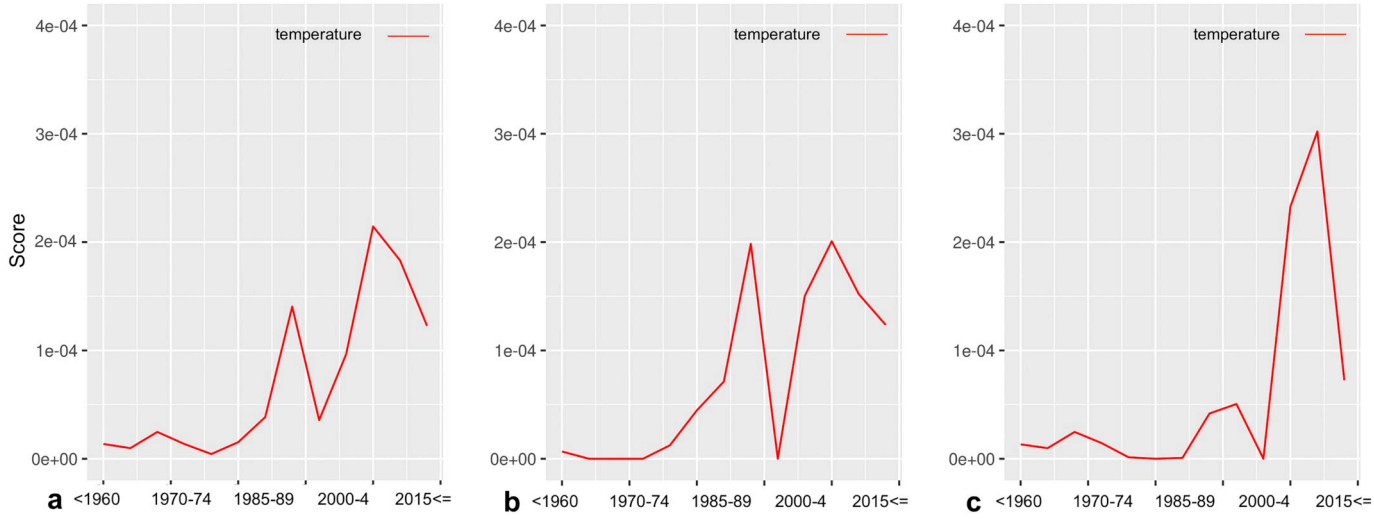

Fig. 10. The term frequency-inverse document frequency scores for temperature over time for all mountain pine beetle documents in the corpus (a), federal agency, (b) and (c) congressional documents.

approach useful for researchers and analysts interested in determining term and topic relevance for ecological issues.

6. Conclusions

Our overall results demonstrate the utility of the semi-automated content analysis approach presented here in providing key insights into relevant topics of policy discussion while also allowing an analysis of important shifts in discourse over time. This method should be seen as a complement, rather than a substitute, for non-automated methods that attempt to more fully interpret the narratives embedded in policy discourse. The semi-automated method we introduce is particularly useful for drawing out broad patterns of association, narrative clusters, and overall trends from very large amounts of text (larger than would generally be analyzable using fully non-automated methods). The interplay between computing capabilities and the judgments of human analysts allows for the production of findings that are both meaningful and quantitatively robust. The case of MPB policy demonstrates the applicability and value of these methods to an issue of particular scientific and managerial importance.

Acknowledgments

This research is based on work supported by the United States National Science Foundation under Grant No. 1414041.

References

Abrams, J., Huber-Stearns, J., Bone, C., Grummon, C., Moseley, C., 2017. Adaptation to a landscape-scale mountain pine beetle empidemic in the era of networked governance: the enduring importance of bureaucratic institutions. Ecol. Soc. 22, 22.
Abrams, J., Huber-Stearns, H., Palmerin, M.L., Bone, C., Nelson, M.F., Bixler, R.P., Moseley, C., 2018. Does policy respond to environmental change events? An analysis of Mountain Pine Beetle outbreaks in the Western United States. Environ. Sci. Pol. 90, 102–109.
Ahonen, H., Heinonen, O., Klemettinen, M., Verkamo, A.I., 1998. Applying data mining techniques for descriptive phrase extraction in digital document collections. In: Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries -ADL'98. IEEE, Santa Barbara, pp. 2–11.
Alghamdi, R., Alfalqi, K., 2015. A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl. 6 (1), 147–153. https://doi.org/10.14569/IJACSA.2015.060121.
Altaweel, M., Bone, C., 2012. Applying content analysis for investigating the reporting of water issues. Comput. Environ. Urban. Syst. 36 (6), 599–613.
Altaweel, M.R., Alessa, L.N., Kliskey, A.D., Bone, C.E., 2010. Monitoring land use: capturing change through an information fusion approach. Sustainability 2 (5), 1182–1203.
Antrop, M., 2001. The language of landscape ecologists and planners. Landsc. Urban Plan. 55 (3), 163–173. https://doi.org/10.1016/S0169-2046(01)00151-7.
Arts, B., 2012. Forests policy analysis and theory use: overview and trends. For. Policy Econ. 16, 7–13.
Becken, S., Stantic, B., Chen, J., Alaei, A.R., Connolly, R.M., 2017. Monitoring the environment and human sentiment on the Great Barrier Reef: assessing the potential of collective sensing. J. Environ. Manag. 203, 87–97.
Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55 (4), 77. https://doi.org/10.1145/2133806.2133826.
Blei, D.M., Carin, L., Dunson, D., 2010. Probabilistic topic models. IEEE Signal Proc. Mag. 27 (6), 55–65.
Bobzein, C., Alstyne, K.V., 2014. Silviculture across large landscapes: back to the future. J. For. 112 (5), 467–473.
Boin, A., Hart, P., McConnell, A., 2009. Crisis exploitation: political and policy impacts of framing contests. J. Eur. Public Policy 16, 81–106. https://doi.org/10.1080/13501760802453221.
Bone, C., Moseley, C., Vinyeta, K., Bixler, R.P., 2016. Employing resilience in the United States Forest Service. Land Use Policy 52, 430–438.
Boyd-Graber, J., Hu, Y., Mimno, D., 2017. Applications of Topic Models. (Boston).
Chau, M., Chen, H., 2008. A machine learning approach to web page filtering using content and structure analysis. Decis. Support. Syst. 44 (2), 482–494.
Cheng, X., Shuai, C., Liu, J., Wang, J., Liu, Y., Li, W., Shuai, J., 2018. Topic modelling of ecology, environment and poverty nexus: an integrated framework. Agric. Ecosyst. Environ. 267, 1–14.
Davis, E.J., Reed, M.G., 2013. Multi-level governance of British Columbia's mountain pine beetle crisis: the roles of memory and identity. Geoforum 47, 32–41.
de Jong, W., Arts, B., Krott, M., 2012. Political theory in forest policy science. For. Policy Econ. 16, 1–6.
de Jong, W., Galloway, G., Katila, P., Pacheco, P., 2017. Forestry discourses and forest based development–an introduction to the special issue. Int. For. Rev. 19, 1–9.
Fifer, N., Orr, S.K., 2003. The influence of problem definitions on environmental policy change: a comparative study of the Yellowstone wildfires. Policy Stud. J. 41 (1), 636–653.
Flannigan, M.D., Stocks, B.J., Wotton, B.M., 2000. Climate change and forest fires. Sci. Total Environ. 262 (3), 221–229.
GitHub, 2019. https://github.com/.
Gong, Y., Xu, W., 2007. Machine Learning for Multimedia Content Analysis. Springer, London.
Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2018. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed. Tools Appl.. https://doi.org/10.1007/s11042-018-6894-4.
Johnstone, J.F., Allen, C.D., Franklin, J.F., Frelich, L.E., Harvey, B.J., Higuera, P.E., Mack, M.C., Meentemeyer, R.K., Metz, M., Perry, G.L.W., Schoennagel, T., Turner, M.G., 2016. Changing disturbance regimes, ecological memory, and forest resilience. Front. Ecol. Environ. 14 (7), 369–378.
Keskitalo, E.C.H., Pettersson, M., Ambjörnsson, E.L., Davis, E.J., 2016. Agenda-setting and framing of policy solutions for forest pests in Canada and Sweden: avoiding beetle outbreaks? For. Policy Econ. 65, 59–68.
Kleinschmit, D., Böcher, M., Giessen, L., 2009. Discourse and expertise in forest and environmental governance — an overview. For. Policy Econ. 11, 309–312. https://doi.org/10.1016/j.forpol.2009.08.001.
Krippendorff, K., 2013. Content Analysis: An Introduction to its Methodology. Sage, London.
Leipold, S., 2014. Creating forests with words—a review of forest-related discourse studies. For. Policy Econ. 40, 12–20.
Liu, Z., Wimberly, M.C., Lamsal, A., Sohl, T.L., Hawbaker, T.J., 2015. Climate change and wildfire risk in an expanding wildland–urban interface: a case study from the Colorado front range corridor. Landsc. Ecol. 30 (10), 1943–1957.
McCarthy, J., 2005. Devolution in the woods: community forestry as hybrid

neoliberalism. Environ. Plan. A 37 (6), 995–1014.
Michalski, R.S., Carbonell, J.G., Mitchell, T.M., 1983. Machine Learning: An Artificial Intelligence Approach. Springer, Berlin, Heidelberg.
Morehouse, B.J., Sonnett, J., 2010. Narratives of wildfire: Coverage in four US newspapers, 1999-2003. Organ. Environ. 23, 379–397.
Murphy, J., Ozik, J., Collier, N., Altaweel, M., Lammers, R., Kliskey, A., Alessa, L., Cason, D., Williams, P., 2014. Water relationships in the U.S. Southwest: characterizing water management networks using natural language processing. Water 6 (6), 1601–1641. https://doi.org/10.3390/w6061601.
Nelson, M., Ciochinna, M., Bone, C., 2016. Assessing spatiotemporal relationships between wildfire and mountain pine beetle disturbances across multiple time lag. Ecospheres 7 (10), e01482.
O'Callaghan, D., Greene, D., Carthy, J., Cunningham, P., 2015. An analysis of the coherence of descriptors in topic modeling. Expert Syst. Appl. 42 (13), 5645–5657.
Parks, S.A., Miller, C., Abatzoglou, J.T., Holsinger, L.M., Parisien, M.A., Dobrowski, S.Z., 2016. How will climate change affect wildland fire severity in the western US? Environ. Res. Lett. 11 (3), 035002.
Pascoe, C.L., Barjat, H., Lawrence, B.N., Tourte, G.J.L., Murray-Rust, P., Hawizy, L., 2016. The PIMMS project and natural language processing for climate science. In: Tonkin, E.L., Tourte, G.J.L. (Eds.), Working with Text: Tools, Techniques and Approaches for Text Mining, pp. 247–269 Oxford.
Polasky, S., Carpenter, S.R., Folke, C., Keeler, B., 2011. Decision-making under great uncertainty: environmental management in an era of global change. Trends Ecol. Evol. 26 (8), 398–404.
Polatkan, A.C., Nieselt, K., 2013. Collact Me: Conceptual Framework for Extracting Domain-Specific Content from Twitter. In: 2013 International Conference on Cloud and Green Computing. IEEE, Karlsruhe, Germany, pp. 313–320. https://doi.org/10.1109/CGC.2013.56.
Popescu, A.-M., Pennacchiotti, M., Paranjpe, D., 2011. Extracting events and event descriptions from twitter. In: In: Proceedings of the 20th International Conference Companion on World Wide Web - WWW '11. ACM Press, Hyderabad, India. https://doi.org/10.1145/1963192.1963246.
Prentice, E.W., Qin, H., Flint, C.G., 2018. Mountain Pine Beetles and ecological imaginaries: The social construction of Forest insect disturbance. In: Urquhart, J., Marzano, M., Potter, C. (Eds.), The Human Dimensions of Forest and Tree Health: Global Perspectives. Springer International Publishing, Cham, Switzerland, pp. 77–107. https://doi.org/10.1007/978-3-319-76956-1_4.
Raffa, K.F., Aukema, B.H., Bentz, B.J., Carroll, A.L., Hicke, J.A., Turner, M.G., Romme, W.H., 2008. Cross-scale drivers of natural disturbances prone to anthropogenic amplification: the dynamics of bark beetle eruptions. Bioscience 58, 501–517.
Rayner, J., McNutt, K., Wellstead, A., 2013. Dispersed Ccapacity and Wweak Ccoordination: the challenge of climate change adaptation in Canada's forest policy sector. Rev. Policy Res. 30, 66–90.
Rehurek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. University of Malta, Valletta, Malta, pp. 46–50.
Röder, M., Both, A., Hinneburg, A., 2015. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, pp. 399–408 New York.
Rosenberger, D.W., Venette, R.C., Maddox, M.P., Aukema, B.H., 2017. Colonization behaviors of mountain pine beetle on novel hosts: implications for range expansion into northeastern North America. PLoS One 12 (5), e0176269.
Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information. Process. Manag. 24 (5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0.
Seidl, R., Thom, D., Kautz, M., Martin-Benito, D., Peltoniemi, M., Vacchiano, G., Wild, J., Ascoli, D., Petr, M., Honkaniemi, J., Lexer, M., Trotsiuk, V., Mairota, P., Svoboda, M., Fabrika, M., Nagel, T., Lexer, M., Reyer, C.P.O., 2017. Forest disturbances under climate change. Nat. Clim. Chang. 7 (6), 395–402.
Sievert, C., Shirley, K.E., 2014. LDAvis: a method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Stroudsburg, pp. 63–70.
Six, D.L., Biber, E., Long, E., 2014. Management for mountain pine beetle outbreak suppression: does relevant science support current policy? Forests 5, 103–133.
Srinivasa-Desikan, B., 2018. Natural language processing and computational linguistics: a practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt, Mumbai.
Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical Dirichlet processes. J. Am. Stat. Assoc. 101 (476), 1566–1581.
Thom, D., Seidl, R., Steyrer, G., Krehan, H., Formayer, H., 2013. Slow and fast drivers of the natural disturbance regime in central European forest ecosystems. For. Ecol. Manag. 307, 293–302.
Vaughn, J., Cortner, H., 2005. George W. Bush's Healthy Forests: Reframing the Environmental Debate. University Press of Colorado, Boulder, CO.
Veltri, G.A., Atanasova, D., 2017. Climate change on Twitter: content, media ecology and information sharing behaviour. Public Underst. Sci. 26 (6), 721–737. https://doi.org/10.1177/0963662515613702.
Walter, L., Radauer, A., Moehrle, M.G., 2017. The beauty of brimstone butterfly: novelty of patents identified by near environment analysis based on text mining. Scientometrics 111 (1), 103–115.
Yau, C.-K., Porter, A., Newman, N., Suominen, A., 2014. Clustering scientific documents with topic modeling. Scientometrics 100 (3), 767–786.
Zhang, Y., Jin, R., Zhou, Z.-H., 2010. Understanding bag-of-words model: a statistical framework. Int. J. Mach. Learn. Cyb. 1 (1–4), 43–52.